Journal of Information Systems and Technology Management – Jistem USP
Vol. 15, 2018, e201815003
ISSN online: 1807-1775
DOI: 10.4301/S1807-1775201815003

RESULTADOS OBTENIDOS EN UN PROCESO DE MINERÍA DE DATOS APLICADO A UNA BASE DE DATOS QUE CONTIENE INFORMACIÓN BIBLIOGRÁFICA REFERIDA A CUATRO SEGMENTOS DE LA CIENCIA

RESULTS OBTAINED IN A DATA MINING PROCESS APPLIED TO A DATABASE CONTAINING BIBLIOGRAPHIC INFORMATION CONCERNING FOUR SEGMENTS OF SCIENCE

E.M. Ruiz Lobaina https://orcid.org/0000-0003-2932-0182
Instituto de Información Científica y Tecnológica, Havana, Cuba

C. P. Romero Suárez https://orcid.org/0000-0003-4640-3502
Instituto Superior de Tecnología y Ciencias Aplicadas, Havana, Cuba

RESUMEN

Este trabajo tiene como objetivo mejorar la calidad de la información que pertenece a la base de datos CubaCiencia, del Instituto de Información Científico y Tecnológico. Esta base de datos tiene información bibliográfica referida a cuatro segmentos de la ciencia y es la base de datos principal del Sistema de Gestión Bibliotecario. La metodología aplicada estuvo basada en los Árboles de Decisión, la Matriz de Correlación, el Scatter Plot 3D, etc., que son técnicas utilizadas por la minería de datos, para el estudio de grandes volúmenes de información. Los resultados alcanzados no solo permitieron mejorar la información de la base de datos, sino que también aportaron patrones verdaderamente útiles en la solución de los objetivos propuestos.

Palabras Claves: Minería de Datos, Minería de Textos, Bibliominería, Descubrimiento de conocimientos, Bibliotecas especializadas.

ABSTRACT

The objective of this work is to improve the quality of the information that belongs to the database CubaCiencia, of the Institute of Scientific and Technological Information. This database has bibliographic information referring to four segments of science and is the main database of the Library Management System. The applied methodology was based on the Decision Trees, the Correlation Matrix, the 3D Scatter Plot, etc., which are techniques used by data mining, for the study of large volumes of information. The results achieved not only made it possible to improve the information in the database, but also provided truly useful patterns in the solution of the proposed objectives.

Keywords: Data Mining, Text Mining, Bibliomining, Knowledge Discovery, Special Libraries.

Manuscript first received: 2014/06/04. Manuscript accepted: 2018/05/03
Address for correspondence:
Esther Marina Ruiz Lobaina, Especialista en Sistemas de Información, Instituto de Información Científica y Tecnológica (IDICT), CITMA, Cuba. E-mail: marina@idict.cu, marinajfr@yahoo.com
Pedro Lázaro Romero Suárez, Profesor Titular, Instituto Superior de Tecnología y Ciencias Aplicadas (INSTEC), Cuba.
E-mail: lromerocu@gmail.com

CC BY

INTRODUCCIÓN

La minería de datos es el proceso más revolucionario hasta el momento, que se encarga de la extracción no trivial de patrones ocultos, útiles y que residen de forma implícita en los datos y también la forma más rápida de estudiar grandes volúmenes de información.

Estas dos razones sirvieron de justificación para que estas técnicas de análisis fueran aplicadas dentro de la tesis doctoral **Metodología para los estudios de Datos Bibliográficos con el empleo de la Minería de Datos** y mientras se conforma el marco teórico, se decidió analizar algunos resultados logrados con las herramientas de minería de datos seleccionada, para determinar que sería lo más apropiado para esta investigación.

Aunque la minería de datos son técnicas de análisis que ya tienen algunos años de explotación en la economía, en los negocios, en la medicina, etc., para la Biblioteca de Ciencia y Técnica del IDICT, es una forma de análisis novedoso, porque por primera vez está siendo aplicada aquí y con sus resultados se ha logrado proponer nuevos productos y servicios que han reanimado el trabajo de la Biblioteca, además que la nueva calidad de la información de la base de datos ha permitido un mejor funcionamiento del Sistema Gestor de Información. (Candás Romero, 2006)

La base de datos que se procesa almacena información científica cubana de las tesis doctorales (T), los premios académicos (PA), las publicaciones seriadas (S), como las revistas científicas y los manuscritos depositados (MD), de cuatro segmentos de las ciencias, como son las Ciencias Agropecuarias, las Ciencias Biomédicas, las Ciencias Técnicas y las Ciencias Sociales. Es una base de datos que ya tiene más de 32 000 registros.

Para lograr estos resultados se ejecutaron varios procesos de minería de datos a la información bibliográfica y se encontraron diferentes patrones, que permitieron hacer mejoras en los productos y servicios bibliotecarios e inclusive dejaron abierta la posibilidad de hacer otras investigaciones futuras. (Rueda-Clausen, Villa-Roel, & Rueda-Clausen, 2005)

BREVE MARCO HISTÓRICO

Desde que los autores Scott Nicholson y Stanton acuñaron en el 2002 que la aplicación de la minería de datos en bibliotecas se denominara bibliominería (*bibliomining*) y también definieron a la bibliominería como *"la combinación de la minería de datos, la bibliometría, la estadística y las herramientas de elaboración de informes y extracción de patrones de comportamiento, basados en los sistemas bibliotecarios" (Nicholson, 2003)*, son muchos los intentos de querer implementar los procesos de minería de datos, dentro de las bibliotecas avanzadas del primer mundo.

Sin embargo en Cuba se ha migrado al software libre buscando ahorro por concepto de compra de softwares, pero los sistemas de gestión bibliotecarios, no incluyen los procesos de minería de datos, por lo tanto, implementar un proceso de minería de datos que permita mejorar la información de la base de datos que trabaja con el sistema de gestión bibliotecario, es imprescindible para obtener el beneficio que aporta el mejoramiento de la información en la base de datos y el valor de los patrones obtenidos con los procesos de minería de datos, porque es la única forma hasta el momento, de reanimar la información guardada en las bases de datos de las bibliotecas y comenzar a brindar una mejor gestión de la información, una mejor gestión del conocimiento, que ayude en el trabajo diario, tanto de los usuarios, como de los propios bibliotecarios.(Herrera Varela, 2006)

CC BY

Es por eso, que este trabajo muestra algunos de los patrones encontrados, y de ellos se analizaron con mayor detalle los patrones logrados con las técnicas de Árboles de Decisiones y la Matriz de Correlación, por ser estos patrones resultados de la clusterización, que aporta conocimiento sobre el comportamiento de la información en la base de datos.

Algunos de estos patrones se muestran en forma textual, otros en forma gráfica y ambas formas permiten hacer una mejor comprensión de los resultados encontrados con las técnicas aplicadas. Ambas salidas ayudan a mejorar servicios y productos, que es uno de los objetivos propuesto.

MATERIALES Y MÉTODOS

Dentro de los materiales empleados se encuentra la base de datos con su información bibliográfica, esta base de datos recoge campos como Nombre del autor, Título de la publicación, Segmento de la ciencia al cual pertenece la publicación, Idioma, Año de publicación, Resumen y seis palabras claves por cada registro (seis metadatos).

La herramienta digital seleccionada para aplicar las técnicas de minería de datos fue Rapid Miner v.5.2, esta es una herramienta libre, que supera al software Weka por la calidad de sus resultados, gráficas y que además ofrece la posibilidad de adjuntarle extensiones, que aumentan las prestaciones digitales que ofrece el software en su forma original.

El método utilizado para la aplicación del proceso de minería de datos consta de cinco procesos fundamentales, las cuales se ordenan de la siguiente manera según su autor (Cabena, 1998):

1. Determinación de los Objetivos.
2. Preparación de datos.
   a. Selección: Identificación de las fuentes de información externas e internas y selección del subconjunto de datos necesario.
   a. Preprocesamiento: estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar.
   a. Transformación de datos: conversión de datos en un modelo analítico.
3. Minería de datos.
   a. Tratamiento automatizado de los datos seleccionados con una combinación apropiada de algoritmos.
4. Análisis de Resultados.
   a. Interpretación de los resultados obtenidos en la etapa anterior, generalmente con la ayuda de una técnica de visualización.
5. Asimilación del conocimiento.
   a. Aplicación del conocimiento descubierto.

Dentro de las técnicas que se utilizaron para encontrar los patrones en la información, están los Arboles de Decisión y la Matriz de Correlación, y dentro de las gráficas que se seleccionaron para

CC BY

mostrar algunos resultados está el Gráfico de Scatter Plot 3D, el Surface 3D y la Desviación. Estos resultados forman parte de una tesis doctoral como ya se mencionó, y hasta aquí se hizo un corte para determinar que ofrecen estos patrones encontrados.

RESULTADOS OBTENIDOS

Después de aplicar las diferentes técnicas de minería de datos, uno de los resultados más interesante logrados, son las salidas de los Árboles de Decisión, porque utilizan las técnicas de Clasificación, además de ser procesos de autoaprendizaje (Madrid, 2009), razón del porque cada resultado es diferente al anterior, también crea una salida en forma de tabla, con todos los metadatos que entraron al análisis.

Los Árboles de Decisión se consideran dentro de los procesos no supervisados (Gutiérrez Rodríguez, 2012), pero el algoritmo requiere conocer a la variable independiente, en este caso la variable fue el Segmento, para a partir de ahí comenzar con la clusterización de las palabras claves y el aprendizaje, por eso se aprecia en la gráfica los cuatros segmentos de la ciencia, cada uno como una rama del árbol. Para este proceso se tomó una muestra de 10 491 registros. Ver Fig. I

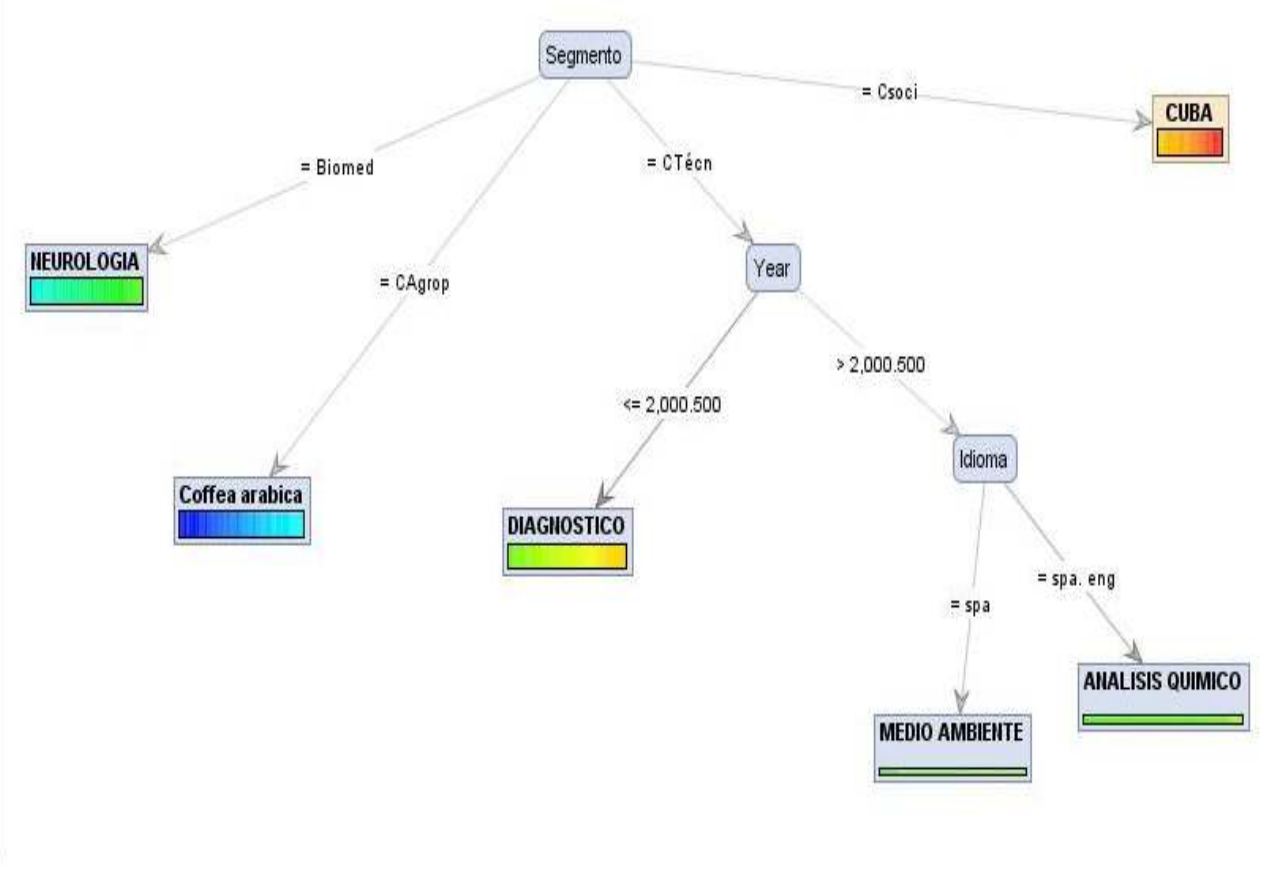


Figura I. Árbol de Decisión logrado con la herramienta Rapid Miner v5.2

Esta gráfica muestra una distribución muy interesante. El segmento Biomed muestra una sola palabra clave (Neurología), el segmento CAgrop igualmente muestra una sola palabra clave, que en

CC BY

este caso es Coffea arábica y Csoci también muestra una sola palabra clave que es Cuba. Estas son las palabras claves con mayor peso en la primera ejecución, sin embargo, para el segmento de CTécn existe un desglose diferente, primero por años (Year) y después por Idioma.

Detectados los errores dentro de la información, se procedió a una segunda revisión y limpieza de la información de la base de datos y nuevamente fue ejecutado el mismo algoritmo, con la misma base de datos, la misma cantidad de registros y tomando nuevamente el mismo juego de variables, donde se declara otra vez que el Segmento es la variable independiente para el proceso de minería de datos y el nuevo resultado emitido, muestra una nueva distribución de la información por Segmento con más detalles, sobre los diferentes metadatos de mayor peso dentro de la base de datos.

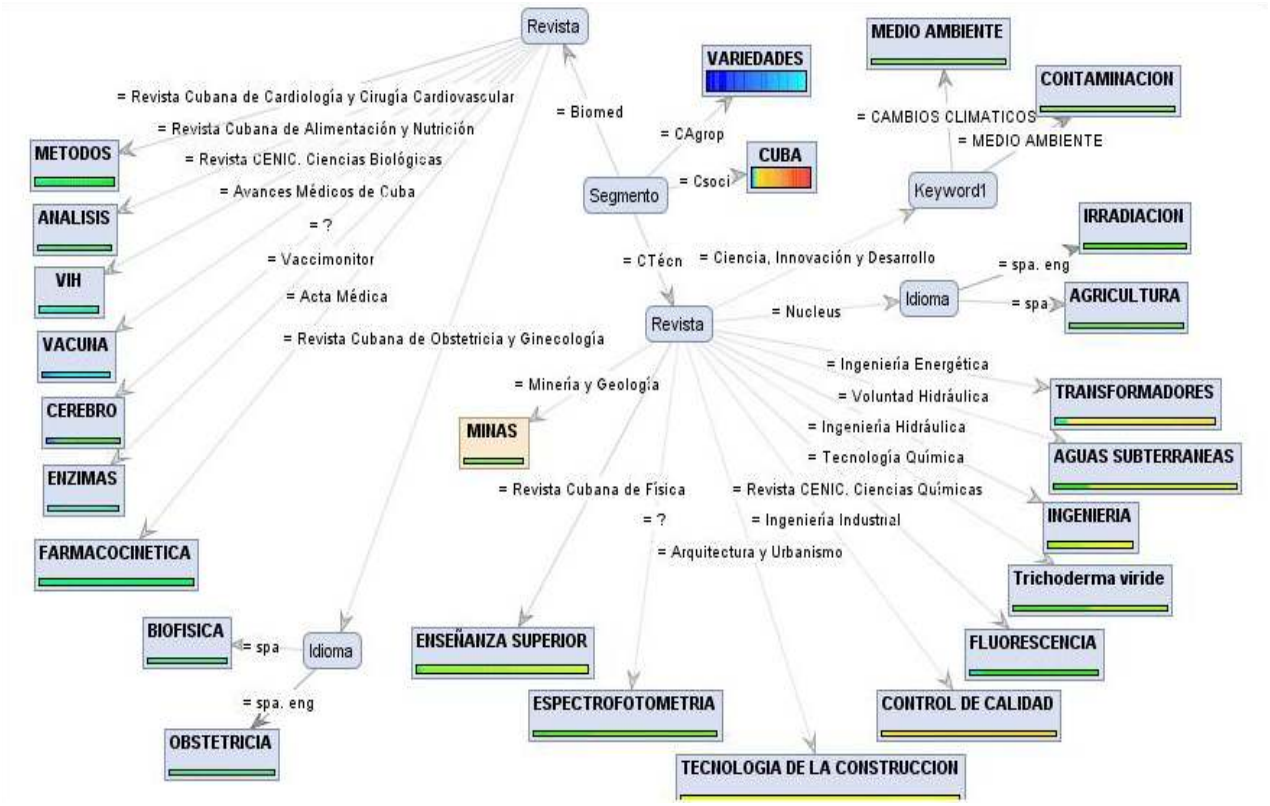


Figura II. Nuevo Árbol de Decisión logrado con Rapid Miner v5.2

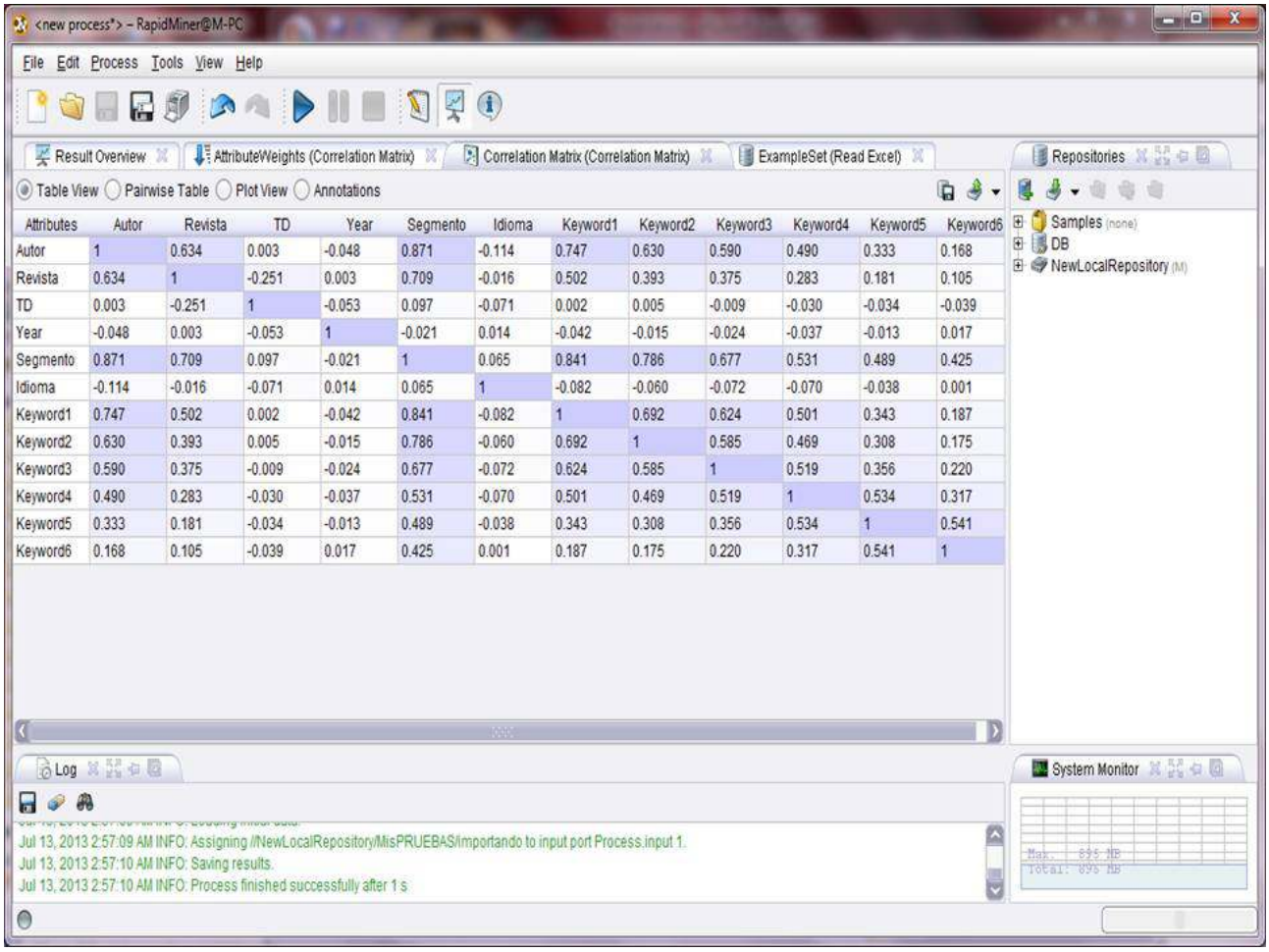
Haciendo una simple comparación de estas dos gráficas (Fig. I y Fig. II), se puede detectar que el Segmento de Csoci, muestra la palabra clave 'Cuba' igual que la Fig. I, mientras que en el caso del Segmento de CAgro se observa que la palabra clave ahora es 'Variedades', a diferencia de la Fig. I, que contenía 'Coffea arábica', sin embargo para los Segmentos de Biomed y CTécn encontramos que la Fig. II despliega diferentes revistas que pertenecen a esos Segmentos, además que hay revistas que también desglosa por Idioma y palabras claves (Keywords).

Estos resultados se lograron con solo mejorar la base de datos en cuanto a la calidad de la información y ejecutar nuevamente el proceso, demostrando de esta manera que una de las fases determinante en los resultados que se obtienen, dependen directamente del pre procesamiento que han tenido los datos, antes de someterse al proceso de minería.

CC BY

Con todos estos patrones logrados a través de los Árboles de Decisión, se puede conocer por cada Segmento las temáticas más investigadas, es decir el comportamiento que están teniendo las investigaciones en el país, y también demuestra que la utilización de la palabra 'Cuba' como el contenido de un metadato es una incorrecta asignación, porque desperdicia la posibilidad de asignar la temática relacionada a ese trabajo de investigación.

Otro de los resultados obtenidos fue la Matriz de Correlación, que ha confirmado relaciones muy interesantes entre variables. El tipo de resultado que muestra la Matriz de Correlación tiene su interpretación basado en el concepto, que mientras más cercano esté el valor a 1, la relación entre variables es más fuerte. Evidentemente la relación más fuerte es la de una variable con ella misma, como se ve en la gráfica. Ver Fig. III



Attributes	Autor	Revista	TD	Year	Segmento	Idioma	Keyword1	Keyword2	Keyword3	Keyword4	Keyword5	Keyword6
Autor	1	0.634	0.003	-0.048	0.871	-0.114	0.747	0.630	0.590	0.490	0.333	0.168
Revista	0.634	1	-0.251	0.003	0.709	-0.016	0.502	0.393	0.375	0.283	0.181	0.105
TD	0.003	-0.251	1	-0.053	0.097	-0.071	0.002	0.005	-0.009	-0.030	-0.034	-0.039
Year	-0.048	0.003	-0.053	1	-0.021	0.014	-0.042	-0.015	-0.024	-0.037	-0.013	0.017
Segmento	0.871	0.709	0.097	-0.021	1	0.065	0.841	0.786	0.677	0.531	0.489	0.425
Idioma	-0.114	-0.016	-0.071	0.014	0.065	1	-0.082	-0.060	-0.072	-0.070	-0.038	0.001
Keyword1	0.747	0.502	0.002	-0.042	0.841	-0.082	1	0.692	0.624	0.501	0.343	0.187
Keyword2	0.630	0.393	0.005	-0.015	0.786	-0.060	0.692	1	0.585	0.469	0.308	0.175
Keyword3	0.590	0.375	-0.009	-0.024	0.677	-0.072	0.624	0.585	1	0.519	0.356	0.220
Keyword4	0.490	0.283	-0.030	-0.037	0.531	-0.070	0.501	0.469	0.519	1	0.534	0.317
Keyword5	0.333	0.181	-0.034	-0.013	0.489	-0.038	0.343	0.308	0.356	0.534	1	0.541
Keyword6	0.168	0.105	-0.039	0.017	0.425	0.001	0.187	0.175	0.220	0.317	0.541	1

Figura III. Matriz de Correlación lograda con la Herramienta Rapid Miner v5.2

Haciendo un breve análisis de las variables de la matriz tenemos que:

1. La variable Autor tiene una relación con las variables Revista, Segmento y con las tres primeras palabras claves (Keyword1, Keyword2 y Keyword3) utilizadas en cada artículo, estando por encima del 0.5, pero pierde relación con el resto de las variables.
2. La variable Revista tiene una relación con las variables Segmento y Keyword1 por encima del 0.5, no así con el resto de las Keywords.
3. Las variables TD, Year e Idioma, son variables que tienen una relación débil con las demás.

CC BY

Este tipo de patrones sirve para hacer estudios sobre la relación que existe entre los campos de la base de datos. En el caso de las palabras claves (Keyword4, Keyword5 y Keyword6) están confirmando la perdida de relación con el contenido de la información, de igual forma que sucedió con los Árboles de Decisión con la Keyword que contiene la palabra Cuba. Se debe tener en cuenta que si de los seis metadatos que tiene cada registro, solo los tres primeros metadatos mantienen una relación fuerte (≥ a 0.5), entre sus metadatos y el tema de los trabajos presentados, mientras que los otros tres metadatos restantes tienen una relación débil, los sistemas de gestión de información no podrán hacer una correcta recuperación de esos registros. Se están desperdiciando tres metadatos para referenciar correctamente cada registro de la base de datos.

Otro inconveniente que esto provoca es sobre la vigilancia tecnológica, estos problemas de mala recuperación también estarían reflejados en este proceso, porque el grado de confiabilidad de la recuperación de la información no es buena.

Entre las salidas que ofrece el software Rapid Miner v 5.2 están los diferentes gráficos, entre ellos se encuentran el gráfico Scatter Plot 3D y Surface 3D, que son utilizados para hacer una representación visual de la matriz de correlación sobre un eje de coordenadas de tres dimensiones. Ver Fig. IV**,** donde se muestran los valores representados en el espacio tridimensional.

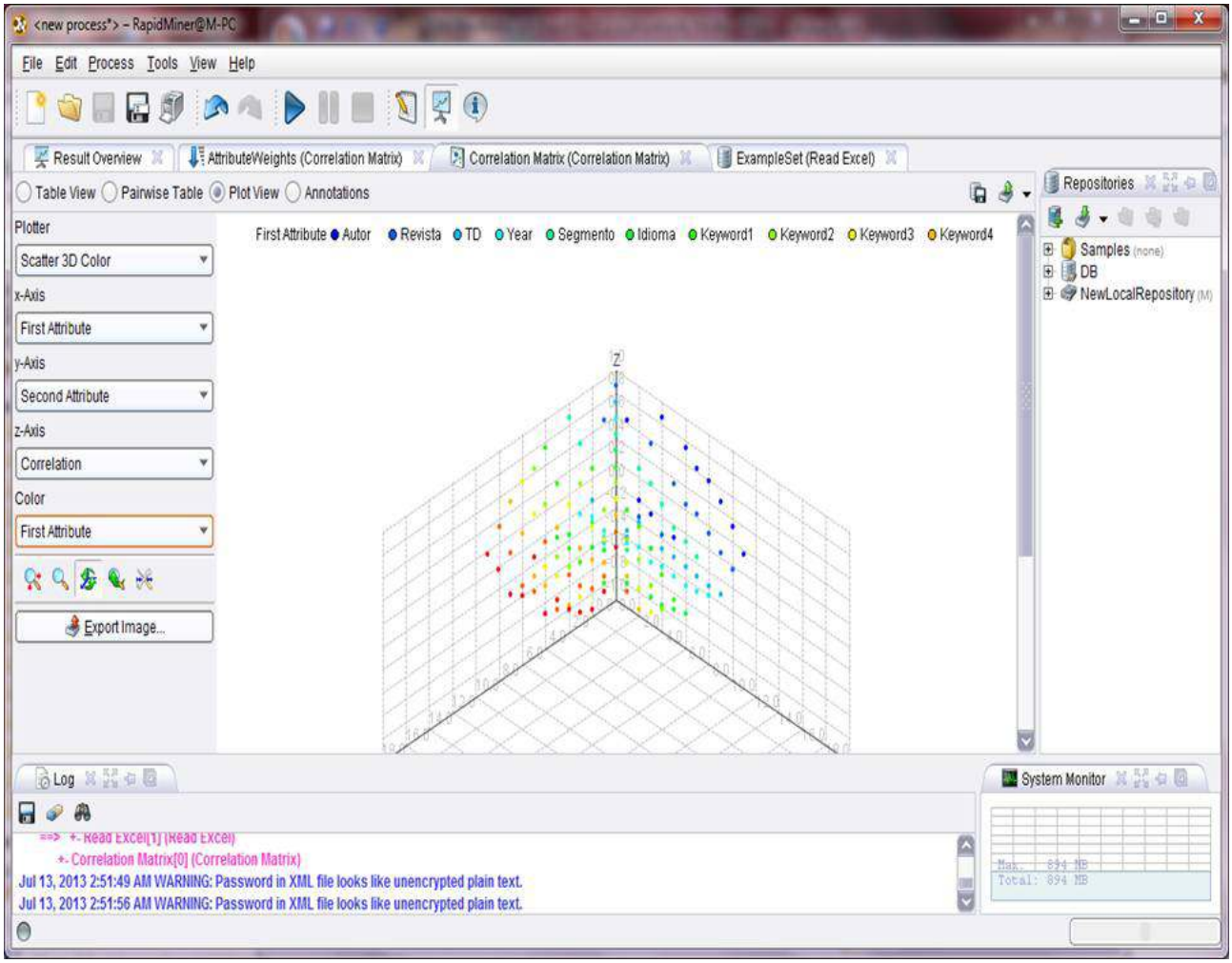


Figura IV. Scatter Plot 3D logrado con la herramienta Rapid Miner v5.2

CC BY

La interpretación de este tipo de gráfica tridimensional consiste en que muestran valores de un conjunto de datos basándose en dos variables, en estos casos el eje X representa la variable independiente o parámetro de control, el eje Y puede representar una variable dependiente o independiente y la coordenada Z muestra el grado de relación que existe entre las variables, que no es otra cosa que el grado de correlación entre las variables, quedando entonces que la representación que suma cada par de valores como las coordenadas de un punto, conforman un conjunto de puntos que se conoce como la nube de puntos o diagrama de dispersión. (Escudero Maximiliano, 2013)

En este gráfico en particular el primer atributo es el Autor por el eje de la X y el segundo atributo es la Revista por el eje de las Y. Cada campo de la base de datos está representado por un color, estas imágenes se pueden ampliar y rotar para ver con mayor claridad la correlación entre variables**.** Dicho de otra forma, la Fig. IV es la representación gráfica de la matriz de Correlación de la Fig. III

Este otro gráfico debajo, llamado Surface 3D, Fig. V, es otra forma de representación tridimensional y al igual que en la Fig. IV plotea la Correlación que existe entre las variables, pero en este caso une los puntos con líneas creando una superficie que da forma a una figura.

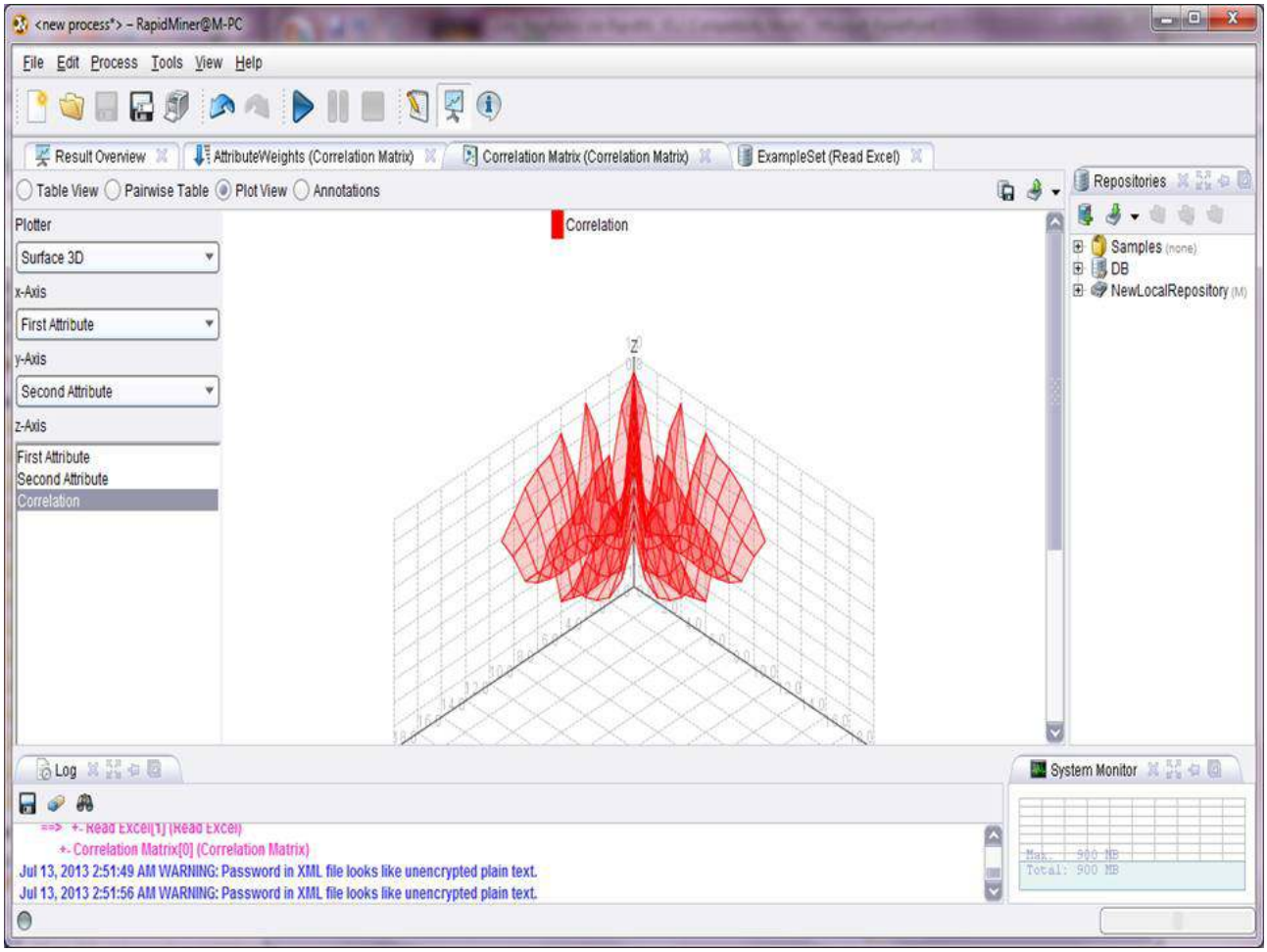


Figura V. Surface 3D logrado con la herramienta Rapid Miner v5.2

En ambas imágenes, Fig. IV y Fig. V, se puede observar que los picos en el diseño de la imagen, se relacionan a los valores más altos de la matriz, al igual que la simetría de la imagen, que corresponde con la simetría de la matriz de correlación Fig. III.

CC BY

En la Fig. VI debajo, se puede observar la representación gráfica de la desviación estándar que tiene el campo Revistas frente al resto de la información de la base de datos. La desviación estándar se calcula como raíz cuadrada de la varianza y se interpreta como la dispersión promedio que hay entre los diferentes valores de la variable respecto de la media aritmética. (Suárez-Ibujes, 2008)

Es la medida de dispersión más importante y juntamente con la media aritmética describen a un conjunto de datos. La desviación estándar de un grupo repetido de medidas nos da la precisión de éstas y la precisión es uno de los valores que se toma en cuenta, para determinar si el modelo escogido es el apropiado. En este caso el modelo es el correcto, porque las medidas de dispersión de todas las variables se mantienen sobre la media, lo que permite suponer que el grado de relación que tiene la información que se encuentra dentro de los campos también está correcta. (Yat Pop, 2008)

La representación visual de la medida de desviación estándar de los datos se puede ver en la Fig. VI

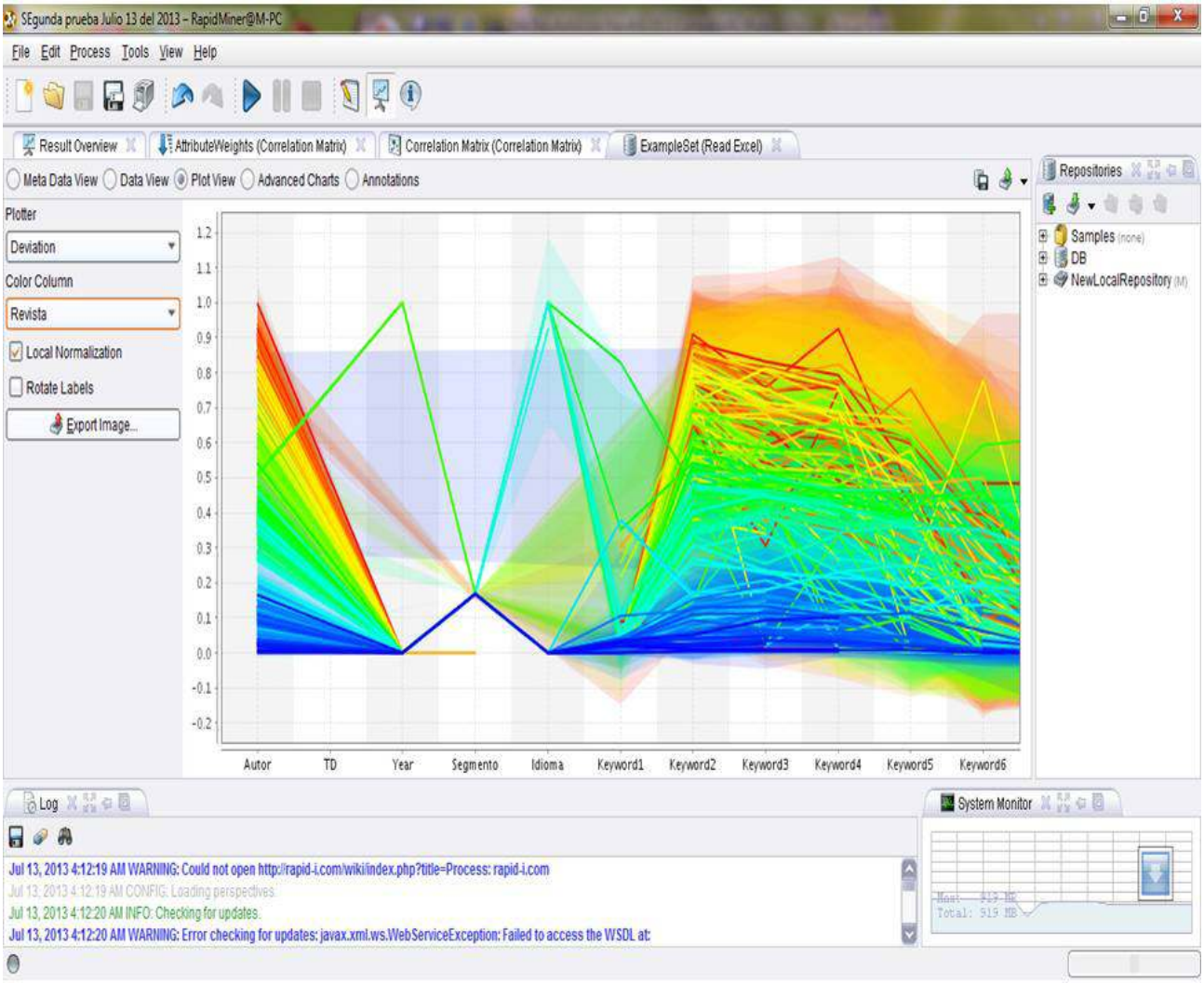


Figura VI. Desviación lograda con la herramienta Rapid Miner v5.2

Es decir, se puede observar que tomando el campo Revista como referencia, existe una gran dispersión de todos los campos restantes. En especial a partir de la Keyword2 a la Keyword4, la dispersión aumenta mientras que ya en la Keyword5, la dispersión comienza a disminuir.

CC BY

No obstante, la dispersión de su valor se mantiene en torno a la media, y por tanto el contenido de esas variables guardan relación directa con el tipo de artículo.

DISCUSIÓN

Con este estudio se planteó el interés de aplicar las técnicas de minería de datos a la información bibliográfica, para lograr los objetivos de:

1. mejorar la calidad de la información de la base de datos, para mejorar el funcionamiento del sistema gestor.
2. encontrar patrones ocultos que sirvan para proponer mejoras en los productos y servicios bibliotecarios.
3. mejorar la gestión de la información y del conocimiento.

Los resultados alcanzados después de la aplicación de las técnicas de minería de datos, demostraron que todos los objetivos propuestos se lograron. Es decir, los diferentes patrones encontrados a través de los Árboles de Decisión, la Matriz de Correlación, la visualización gráfica de la Correlación a través del Scatter 3D Plot, el Surface 3D y la Desviación Standard, permitieron mostrar que:

Durante el desarrollo del proceso de minería de datos que abarca los aspectos relacionados a la Preparación de datos, quedó demostrado que en la selección del subconjunto de datos y la selección de los campos seleccionados para aplicar la minería de datos fue la correcta.

Durante el Preprocesamiento de la información, se logró mejorar la calidad de la información que tiene la base de datos, a través de la estandarización de la información, característica que permite que la información pueda ser utilizada, por cualquier otro sistema de gestión de información.

Los Árboles de Decisión mostraron las temáticas que más se están investigando a nivel nacional, ordenada por los campos Segmentos y Revistas. Con esta información se propuso la instalación de un repositorio que sirve de apoyo a un servicio de vigilancia tecnológica, que permite conocer el avance que tienen las investigaciones a nivel nacional en comparación con la de otros países.

CONCLUSIONES

Se cumplió con el mejoramiento de la calidad de la información de la base de datos y esto permitió un mejor funcionamiento del Sistema de Gestión Bibliotecario.

Se cumplió con el objetivo de proponer nuevos productos y servicios para la Biblioteca, haciendo uso de los patrones encontrados a través de la minería de datos.

Se creó un repositorio con los resultados de los patrones de la minería de datos, que es usada para estudios de vigilancia y prospección.

CC BY

REFERENCIAS BIBLIOGRÁFICAS

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. Zanasi, A. . (1998). Discovering Data Mining: From Concept to Implementation. from http://www.zanasi-alessandro.eu/publications/cabena-p-hadjinian-p-stadler-r-verhees-j-zanasi-a-1998-discovering-data-mining-from-concept-to-implementation/

Candás Romero, J. (2006). Minería de datos en bibliotecas: bibliominería. from http://www.ub.edu/bid/17canda2.htm

Escudero Maximiliano, J., Lujan Ganuza, M., Wilberger, D., Martig, Sergio R. (2013). Scatter Plot 3D. from http://sedici.unlp.edu.ar/bitstream/handle/10915/20366/Documento_completo.pdf?sequence=1

Gutiérrez Rodríguez, A. E., García Borroto, M. & Martínez Trinidad, J.F (2012). Algoritmo de agrupamiento basado en patrones utilizando árboles de decisión no supervisados. from http://3c.inaoep.mx/portalfiles/CCC-12-002.pdf

Herrera Varela, R. (2006). Bibliomining: minería de datos y descubrimiento de conocimiento en bases de datos aplicados al ámbito bibliotecario. from http://lemi.uc3m.es/est/forinf@/index.php/Forinfa/article/view/122/127

Madrid, U. C. I. d. (2009). Análisis de Cluster y Arboles de Clasificación. from http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema6dm.pdf

Nicholson, S. (2003). Bibliomining for automated collection development in a digital library setting: Using data mining to discover web-based scholarly research works. Journal of the American Society for Information Science and Technology,54(12). from http://bibliomining.com/nicholson/asisdiss.html

Rueda-Clausen, G. C. F., Villa-Roel, G. C., & Rueda-Clausen, P. C. E. (2005). Indicadores bibliométricos: origen, aplicación, contradicción y nuevas propuestas. MedUNAB, Vol 8, No 1. from http://revistas.unab.edu.co/index.php?journal=medunab&page=article&op=view&path[]=208&path[]=191

Suárez-Ibujes, M. O. (2008). Conceptos básicos de Probabilidades y Estadística Inferencial. from https://es.scribd.com/doc/129480693/Conceptos-basicos-de-Probabilidades-y-Estadistica-Inferencial#download

Yat Pop, O. (2008). Regresión y Correlación. from http://oscarmanuelyatpop.blogspot.com/2008/06/regresion-y-correlacion.html

CC BY